

PacI		Bse RI	Bsp 120		Bbs I	Eco RI	Bam HI
↓		↓	↓		↓	↓	↓
aattgttaattaaggatgagctcactcctcgggcccgcataagtcttcgaattcg							
caattaattcctactcgagtgaggagcccgggcggtattcagaagcttaagcctag							

## Formula II

Separately, the oligonucleotide of Formula I and forward and reverse primers (SEQ ID NO: 2) and SEQ ID NO: 3) are synthesized using a conventional DNA synthesizer, e.g. PE Applied Biosystems (Foster City, CA) model 392. The oligonucleotide of Formula I is a mixture containing a repertoire of 64 two-word oligonucleotide tag precursors. The four-nucleotide words of Table I are employed. After amplification by PCR the amplification product is digested with Bbs I to give the following two products (SEQ ID NO: 28 and 29):

... gaagacga word-word-gg ...  
 ... cttctgct-word word-cc ...

The products are re-ligated, amplified by PCR, and digested with Bbv I to give the following two products (SEQ ID NO: 30 and 31):

... gaagacga-word word-gg ...  
 ... cttctgct-word-word cc ...

The products are again re-ligated and amplified by PCR. By this sequence of cleavages and relations, any words consisting of failure sequences are selected against by the ligation event, i.e. words with failure sequences will not religate in the mixture, and thus, will not be amplified. The final product is digested with Pst I and Hind III and inserted into a Pst I/Hind III-digested pUC19 To give the following construct (SEQ ID NO: 5):

Pst I	Bse RI	Bbs I	Bsp 120	Hind III
↓	↓	↓	↓	↓
...cgacctgcagaggagatgaagacga-wordword-gggcccaatgctgcaagcttggcg...				
...gctggacgtctcctctacttctgct-wordword-cccgggttacgacgttcgaaccgc...				
				↑
				Bbv I

where Pst I, Bse RI, Bbs I, Bsp 120, and Bbv I, correspond to  $r_4$ ,  $r_5$ ,  $r_6$ ,  $r_7$ , and  $r_8$  of Figure 2, respectively. After amplification in a suitable host, the plasmid is isolated and cleaved with Pst I and Bbs I to give an opened vector with the following upstream and downstream (SEQ ID NO: 6 and 32) ends:

...cgacctgca	wordword-gggcccaatgctgcaagcttggcg...
...gctgg	word-cccgggttacgacgttcgaaccgc...

02 Separately, a portion of the amplified oligonucleotide of Formula I is digested with Pst I and Bbv I to give the following fragment (SEQ ID NO: 7):

gaggagatgaagacga-word  
acgtctcctctacttctgct-wordword

This fragment is inserted into the above vector opened by digestion with Bbs I and Pst I to give the following construct (SEQ ID NO: 8):

...gcagaggagatgaagacga-wordwordword-gggcccaatgctgcaagcttggcg...
...cgtctcctctacttctgct-wordwordword-cccgggttacgacgttcgaaccgc...

which contains an oligonucleotide tag precursor of three words. The steps of cleaving, inserting, and amplification are repeated until a construct containing eight words is obtained. Preferably, at each step, reactants, e.g. vectors and/or inserts, are provided in amounts that are at least ten times the complexity of the reactant. When synthesis is complete, the eight-word construct is cleaved with

Bse RI and Bsp 120 and the following fragment containing the oligonucleotide tag repertoire is isolated (complement is SEQ ID NO:33):

(word)<sub>8</sub>g  
ct (word)<sub>8</sub>ccccgg

*A2  
done*  
The isolated fragment is then inserted into the Bse RI/Bsp 120 vector of Formula II, which vector is used to transform a suitable host. The construct is ready for inserting polynucleotides, such as cDNAs, into the Eco RI restriction site to form tag-polynucleotide conjugates in accordance with the method of Brenner et al., International patent application pct/us96/09513.

*2* Please replace the paragraph starting at page 17, line 30 with the following:

After cloning, the population of vectors is divided into two parts, after which the vectors in one part are cleaved with Pst I and Bsg I to give the following fragment mixture (SEQ ID NO:11 and 34):

gttatcggaggagatgaagacgg[word][word]gg  
acgtcaatagcctcctctacttctgcc[word][word]

*A3*  
which is isolated. The vectors in the other part are cleaved with Pst I and Bse RI and the linearized word-containing vectors are isolated. The word-containing fragments are ligated into the linearized vectors to form the following construct (SEQ ID NO: 12):

...ctgcagttatcggaggagatgaagacgg[word][word]gg[word][word]-  
...gacgtcaatagcctcctctacttctgcc[word][word]cc[word][word]-

-gggcccataatatccgtctgcacaagcttacg...  
-ccccgggtatataggcagacgtgttcgaaccgc...

After cloning, the construct is again divided into two parts and the steps are repeated to give the final 8-word repertoire having the form (SEQ ID NO: 35):

.. gaagacgg ([word] [word] gg) <sub>4</sub>gccc ...  
 .. cttctgcc ([word] [word] cc) <sub>4</sub>cggg ...

23  
cont This may then be cleaved with Bse RI and Bsg I and re-cloned into a vector similar to that of Formula II for attachment to polynucleotides.

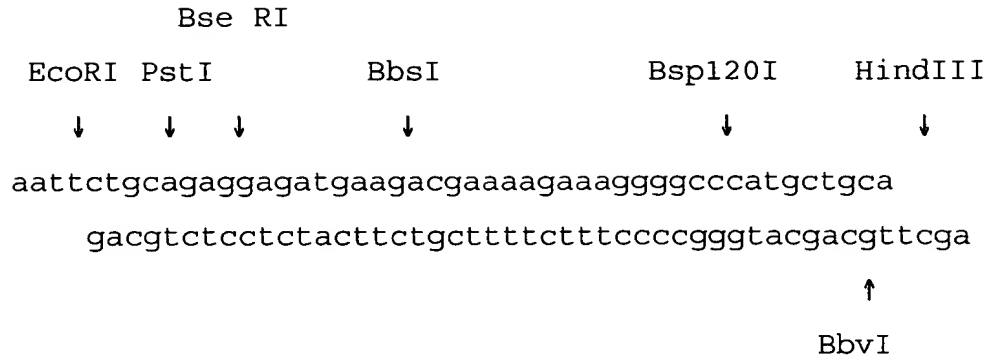
2 Please replace the paragraph starting at page 18, line 36 with the following:

pUC19 was digested to completion with Sap I and Eco RI using the manufacturer's protocol and the large fragment was isolated. All restriction endonucleases unless otherwise noted were purchased from New England Biolabs (Beverly, MA). The small Sap I-Eco RI fragment was removed to eliminate the  $\beta$ -gal promoter sequence, which was found to skew the representation of some combinations of words in the final library. The following adaptor (SEQ ID NO: 13 and 36) was ligated to the isolated large fragment in a conventional ligation reaction to give plasmid pUCSE as a ligation product.

24

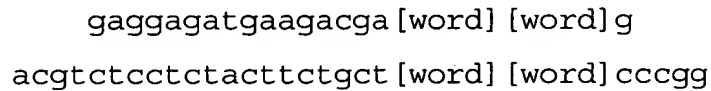
Eco RI	Pst I	Eco RV	Hind III
↓	↓	↓	↓
aattctagactgcagttgatatcttaagctt			
gatctgacgtcaactatagaattcgaacga			

A bacterial host was transformed by the ligation product using electroporation, after which the transformed bacteria were plated, a clone was selected, and the insert of its plasmid was sequenced for confirmation. pUCSE isolated from the clone was then digested with Eco RI and Hind III using the manufacturer's protocol and the large fragment was isolated. The following adaptor (SEQ ID NO: 14 and 37) was ligated to the large fragment to give plasmid pUCSE-D1 which contained the first di-word (underlined).



### Formula I

Further plasmids, pUCSE-D2 through pUCSE-D64, containing di-words were separately constructed from pUCSE-D1 by digesting it with Pst I and Bsp 120 I and separately ligating the following adaptors (SEQ ID NO: 15) to the large fragment.



### Formula II

The words of the top strand were selected from the following minimally cross-hybridizing set: gatt, tgat, taga, ttg, gtaa, agta, atgt, and aaag. After cloning and isolation, the inserts of the vectors were sequenced to confirm the identities of the di-words.